

INFORMATION COLLECTION DEVICE AND INFORMATION COLLECTION PROGRAM

Patent number: JP2003271610
Publication date: 2003-09-26
Inventor: YAMADA TAKESHI; KASE NAOKI
Applicant: TOKYO SHIBAURA ELECTRIC CO
Classification:
- international: **G06F17/30; G06F17/30;** (IPC1-7): G06F17/30
- european:
Application number: JP20020075412 20020319
Priority number(s): JP20020075412 20020319

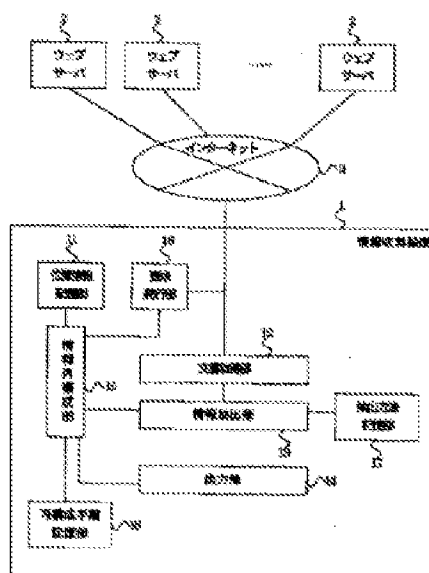
Report a data error here

Abstract of JP2003271610

PROBLEM TO BE SOLVED: To moderate the risk accompanied by information collection with disclosed information on the Internet as an information source and to perform a reliable information collection.

SOLUTION: A plurality of sites for providing intended information is selected from sites disclosed on the Internet. More probable information is selected from respective pieces of information provided by the plurality of sites, reconfigured and outputted.

COPYRIGHT: (C)2003,JPO



.....
Data supplied from the **esp@cenet** database - Worldwide

(11)特許出願公開番号
特開2003-271610
(P2003-271610A)

(51)Int.Cl. ⁷	識別記号	F I	データベース ⁸ (参考)
G 0 6 F 17/30	1 8 0	G 0 6 F 17/30	1 8 0 Z 5 B 0 7 J
	1 1 0		1 1 0 C

【特許請求の範囲】

【請求項1】 ネットワークを通じて電子的に公開されている情報をアクセスする情報収集装置であって、収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報を予め記憶しておく位置情報記憶手段と、一度の取得要求に応じて前記複数の位置情報を用いて実質的に同一の目的とする情報を複数の所在から取得しようとする情報取得手段とを具備したことを特徴とする情報収集装置。

【請求項2】 前記情報取得手段は、前記位置情報記憶手段が記憶したある位置情報から目的とする情報が取得できなかった場合には、前記位置情報記憶手段が記憶する他の位置情報を用いて、該他の位置情報が指し示す別の目的とする情報から収集対象に関する情報を取得しようとすることを特徴とする請求項1に記載の情報収集装置。

【請求項3】 取得した複数の目的とする情報から収集対象情報に関する情報要素を生成する情報生成手段をさらに具備し、前記情報生成手段は、複数の目的とする情報に基づいて得られる情報要素の内から出力として採用するものを、予め与えられた手順に従って選択する選択手段を有するものであることを特徴とする請求項1乃至2に記載の情報収集装置。

【請求項4】 前記情報取得手段は、一度の取得要求に応じて、複数の収集対象情報に関する情報をまとめて取得することを特徴とする、請求項1乃至3に記載の情報収集装置。

【請求項5】 前記情報生成手段によって選択された前記情報要素を、予め指定した出力様式に従って構成し直す情報構成手段をさらに具備することを特徴とする、請求項4に記載の情報収集装置。

【請求項6】 前記選択手段は、複数の前記情報要素の内から出力として採用するものを選択するとき、位置情報によって示されるそれぞれの情報提供サーバの稼動状況と、位置情報によって示されるそれぞれの目的とする情報の、統計的手法に基づく取得対象情報に関する信頼度との少なくともどちらかの、取得状況によって動的に変更される、あるいは予め与えられる情報に基づいて選択が為されることを特徴とする請求項3乃至5に記載の情報収集装置。

【請求項7】 ネットワークを通じて電子的に公開されている情報をアクセスする情報収集プログラムであって、一度の取得要求に応じて、収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報から、複数の位置情報を選択する第1のステップ

と、

前記第1のステップで選択した目的とする情報についての複数の位置情報を用いて、収集対象情報について実質的に同一の複数の目的とする情報を取得する第2のステップとを有することを特徴とする情報収集プログラム。

【請求項8】 前記第2のステップで、ある位置情報から目的とする情報が取得できなかった場合には、複数記憶する他の位置情報を用いて、該他の位置情報が指し示す別の目的とする情報から収集対象に関する情報を取得しようとする第3のステップをさらに有することを特徴とする請求項7に記載の情報収集プログラム。

【請求項9】 複数の目的とする情報に基づいて得られる情報要素の内から出力として採用するものを、予め与えられた手順に従って選択し、取得した複数の該目的とする情報から収集対象情報に関する情報要素を生成する第4のステップをさらに有することを特徴とする請求項7乃至8に記載の情報収集プログラム。

【請求項10】 前記第2または第3のステップは、一度の取得要求に応じて、複数の収集対象情報に関する情報をまとめて取得することを特徴とする、請求項7乃至9に記載の情報収集プログラム。

【請求項11】 前記第4のステップによって選択された前記情報要素を、予め指定した出力様式に従って構成し直す第5のステップをさらに有することを特徴とする、請求項10に記載の情報収集プログラム。

【請求項12】 前記第4のステップは、複数の前記情報要素の内から出力として採用するものを選択するとき、位置情報によって示されるそれぞれの情報提供サーバの稼動状況と、位置情報によって示されるそれぞれの目的とする情報の、統計的手法に基づく取得対象情報に関する信頼度との少なくともどちらかの、取得状況によって動的に変更される、あるいは予め与えられる情報に基づいて選択が為されることを特徴とする請求項9乃至11に記載の情報収集プログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、公開されている情報の収集装置及びプログラムであって、特に複数の情報源から情報を収集し目的とする情報あるいは情報要素の取得確率を高めるとともに、信頼性が高い情報を収集する情報収集装置及びプログラムに関する。

【0002】

【従来の技術】インターネットの発達によって電子化されたさまざまな情報を入手することが容易になってきた。昨今では、今までは対価を払って入手していた情報も、インターネットから無償で入手できることが多い。あるいは高価な対価を払って契約する既存の情報提供サービスも、インターネットを介して行なわれることが多くなってきた。インターネット上で得られる情報は、一般にウェブページの形で提供されており、インターネッ

トの利用者はウェブブラウザを使用してこれらの情報にアクセスしている。

【0003】これらウェブページを元にした情報収集には大きく二つの問題がある。一つめはHTML(HyperText Markup Language)形式で提供されるために、そこに含まれる情報の二次利用が困難であることである。HTML形式は画面上に視覚的な情報を投影するには好都合であるが、反面、データが整然と表記されたものにはならない。この問題をある程度克服するものとして特開2000-348061号公報にあるような、文書構造を解析しデータの抽出およびデータの二次利用を行なうための発明がされている。

【0004】もう一つは、ワールドワイドに点在する、情報源となるウェブサイトの全てが必ずしも厳格な管理の元に運営されているわけではないことである。提供する情報の信頼性を保証する契約の元に運営されるウェブサイトは別として、インターネット上の無償で提供される有益なウェブサイトのほとんどは、提供する情報の信頼性を保証していない。これは当該ウェブサイトの運営が保証されているものではなく、また必ずしも常に正しい情報を提供しなければならない義務を負っているものでもないことを意味している。このような実情から、情報源となるそれぞれのウェブサイトはウェブサイト運営者の事情による不意の停止や提供ウェブページの変更、あるいは提供情報に誤りが有り得ることを十分考慮しなければならない。上記実情を鑑みれば、定常的にサービスを行なうシステムの情報源として既述のウェブサイトを使用することは困難であり情報更新のリアルタイム性が確保できないばかりでなく、場合によっては自らのサービスの停止を余儀なくされることさえ有り得る。さらには誤った情報に基づいて情報をサービスし続けてしまうことなども懸念される。このことは当該システムを提供する企業等にとって信用失墜などの極めて重大な問題を引き起こし、またそれによる機会損失を蒙るなど、その悪影響は計り知れない。

【0005】

【発明が解決しようとする課題】インターネット上の公開情報を用いて情報収集を行なう際の上記のような問題を緩和し、インターネット上の公開情報を情報源とした情報の取得率と提供する情報の信頼性の向上を図ることを目的とする。

【0006】

【課題を解決するための手段】第1の発明によれば、ネットワークを通じて電子的に公開されている情報をアクセスする情報収集装置であって、収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報を予め記憶しておく位置情報記憶手段と、一度の取得要求に応じて前記複数の位置情報を用いて実質的に同一の目的とする情報

を複数の所在から取得しようとする情報取得手段とを具備したことを特徴とする情報収集装置が提供される。また、第2の発明によれば、ネットワークを通じて電子的に公開されている情報をアクセスする情報収集プログラムであって、一度の取得要求に応じて、収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報から、複数の位置情報を選択する第1のステップと、前記第1のステップで選択した目的とする情報についての複数の位置情報を用いて、収集対象情報について実質的に同一の複数の目的とする情報を取得する第2のステップとを有することを特徴とする情報収集プログラムが提供される。

【0007】これによって、実質的に同一の情報を複数収集し、継続的な目的とする情報の収集と、より信頼のおける情報の取捨選択とを行なわせることが可能となる。

【0008】

【発明の実施の形態】第1の実施形態における情報取得装置のシステム構成図の一例を図1に示す。情報収集処理装置1と、実質的に相互に同等な情報を内包する半構造化文書、たとえばHTML(HyperText Markup Language)形式のウェブページを電子的手段によって提供するサーバ、たとえばウェブページを提供するウェブサーバ2と、それが複数接続されたインターネット3とが示されている。

【0009】次に情報収集装置1について説明する。情報再構成部15は再構成手順記憶部18に記憶された文書取得手段に従って、少なくとも1つ以上のウェブページを文書取得部12に指示して取得させる。そして各々のウェブページに内在する情報要素を、抽出方法記憶部17に記憶された抽出方法に従って情報抽出部13に抽出させる。このようにして抽出された情報要素は、情報再構成部15によって情報の再構成、たとえばHTMLへの再編集あるいはリスト形式のデータに加工されるなどして、出力部14を介して出力される。

【0010】図2は情報収集装置1の動作のフローの一例を示す図である。情報再構成部15は位置情報記憶部11に記憶されたそれぞれのウェブサーバ2の、たとえばURL(Uniform Resource Locator)を取得する(ステップS1)。URLはサーバがインターネット3における位置と、ウェブページの格納場所を示す位置情報の組を示す規定の形式である。続いて当該URLが示すウェブページを取得すべく、要求発行部16からインターネット3を通じてウェブサーバ2に対して要求する(ステップS2)。要求を受け取ったウェブサーバ2はこれに呼応してURLにて指定されたウェブページを情報収集装置1に対して送り返す。情報収集装置1は、この送り返されてきたウェブページを文書取得部12にて受け取る(ステ

ップS3)。このとき、ウェブサーバ2やインターネット3等で不具合が生じない限り、URLで表現したウェブページが取得できる。

【0011】取得に成功したウェブページはある文書構造（たとえばHTML）にしたがって情報要素（たとえばHTML内の株価情報など）を内包している。各々のウェブページはそれぞれが別個の文書構造をしているのが普通であるから、情報抽出部13はそれぞれの文書構造に応じた予め記憶してあるデータ取得方法を抽出方法記憶部17から取得（ステップS4）し、その取得方法によって出力となる1組の情報要素群を抽出する（ステップS5）。

【0012】情報要素の抽出に際しては、取得したウェブページが抽出方法記憶部17に記憶された抽出方法を検討したときと同じものであれば何ら問題は生じない。場合によっては、たとえばウェブサーバ2の運営者によって当該ウェブページを変更した場合や、ウェブサーバ2の不具合によって指示したのとは別のウェブページが取得される場合も考えられる。どのようなウェブページが取得されたかにもよるが、一部あるいは全部の情報要素の取得に失敗する、もしくは誤った情報要素を抽出してしまう。このことは情報収集という目的において大きな問題となる。しかしながらこのような事態の発生は予測が不可能である上に、外的要因によるものであるため、一つのURLで示されるウェブページに頼る方法では対処が難しい。そこで情報再構成部15は再構成手順記憶部18から予め記憶しておいた再構成手順を取得（ステップS6）し、その手順に従って抽出された複数の重複した情報要素を元に出力となる情報要素群を再構成する（ステップS7）。再構成手順には、各情報要素がどの目的とする情報に含まれているか、重複する情報要素の中で異なるデータが出現したときの対処の方法、そしてどのようなデータであれば信頼できるか等の情報を含んでいる。場合によっては情報再構成部15が再構成手順に記録された方法によって再度ウェブページを取得して、新規に出力となる情報要素群を取得することも考えられる。

【0013】このようにして構成された出力結果は出力部14を介して出力され（ステップS8）、一連の処理を終了する。

【0014】上記のように構成することによって、情報収集という目的における一部のウェブページの取得失敗や一部のウェブページに掲載された誤った情報に対する耐性を備えることが可能となる。このことによって出力する情報に欠損や誤りが生じる可能性を大きく低減させるとともに、情報の信頼性を高めることができる。

【0015】既述した一連の動作は、定期的に繰り返されることによってリアルタイムに変化する情報、たとえば証券金融情報や気象情報などを最新情報に近い状態を維持しつつ常時、情報を提供するサービスにも資するこ

ともできる。

【0016】次からは実際に取得するウェブページの例を示しながら説明する。ここに示す実施例は、ウェブページの作成言語として一般的に利用されているHTMLを例にとっている。本発明では抽出方法記憶部17に記憶する抽出方法を適切に設計することで、ウェブページがXHTMLなどその他の形式で提供された目的とする情報であっても対処可能である。

【0017】図3および図5は、1日のある時点の株価情報をウェブブラウザで提供するために作成されたHTML文書の簡易な例を示している。そして図4および図6は、図3および図5を一般のウェブブラウザによって表示させたときの表示イメージを表している。例では株価情報を提供するウェブサイトが2箇所あり、そのそれぞれが提供するHTML文書が図3および図5であるとしている。当然のこととして、図3および図5に示したHTML文書をウェブブラウザで表示されると、図4および図6のように異なったものとなる。図3のものは、`<TABLE..></TABLE>`で囲まれた最初の領域に株価データが記録されているが、図5では2番目の`<TABLE..></TABLE>`領域に株価データが記録されている。

【0018】このため図3のものと図5のものから株価情報を抽出するには、それぞれ異なる抽出方法を用いて行なわねばならない。これはそれぞれ個別の抽出方法が抽出方法記憶部17に記憶されている必要があることを意味している。以降、説明のために図3のHTML文書を提供するウェブサーバをウェブサーバA、図5の方をウェブサーバBと称する。

【0019】また、ウェブサーバAとウェブサーバB（表示画面はそれぞれ図4と図6）では、共通して保持されている情報と片方にしか存在しない情報とがある。銘柄名、現在値、取引時刻の3点は両方のウェブサーバから得ることができる。またウェブサーバAからは前日終値が、ウェブサーバBからは前日比が提供されるが、これは同一銘柄について見れば相互に補完が可能（前日比＝現在値－前日終値）であるから、相互に取得可能な情報要素と見ることができる。

【0020】一方、ウェブサーバAが提供する始値、高値、安値、およびウェブサーバBが提供する出来高はそれぞれ片方にしか存在せず、また例示したような補完関係にもないため、片方のウェブサーバからのみ取得が可能な情報要素である。

【0021】図7および図8は、それぞれウェブサーバAおよびBのそれぞれから株価情報を抽出するための抽出方法の一例を図示したものである。これらは抽出方法記憶部17に記憶され、情報抽出部13が情報要素を抽出する際に使用される。この例では情報抽出部13は、ウェブサーバAあるいはBから与えられるHTML文書を、図7および図8に示した抽出手順に従って抽出していく。まず図7および図8に示すデータ開始位置から下方

に参照する。銘柄情報開始位置以降の銘柄の各データを、図7および図8の3～7行目に記載された位置のデータを書式に従って抜き出す。必要に応じて計算も行いながら、データ終了位置まで、銘柄ごとに繰り返す。すると情報抽出部13は銘柄名(A社、B社、C社)とそれぞれの現在値、取引時刻、前日終値、前日比を取得できる。株価自体は共通のものであるから、原則としてウェブサーバAとウェブサーバBから得られるこれら共通に提供される情報要素は同一の値となるはずである。

【0022】ここで各ウェブサーバが固有に提供するデータを取得したい場合は、高値、安値、出来高を抽出するための抽出方法を、図7および図8に示す抽出方法に書き加え抽出方法記憶部17に記憶させればよい。さらにウェブサーバAやウェブサーバB以外のウェブサーバCにおいて高値、安値、出来高が提供されるのであれば、このウェブサーバCの該当ウェブページを指し示すURLとそのウェブページについての抽出方法を、位置情報記憶部11と抽出情報記憶部17に加えることにより重複して取得することができるようになる。

【0023】このようにして得られた重複した情報要素は、全て情報再構成部15に送られる。そして情報再構成部15はこれら重複した情報要素から出力となる1組の情報要素群を構成する。以下、取得時の情報要素のさまざまな取得状況において当該情報再構成部15が取り得る、信頼性の高い情報要素群を再構成する再構成手順の事例を示す。

【0024】情報再構成部15が取得する複数の情報要素群は、本来同じ情報となるはずである。しかしながらさまざまな事情や不具合によって期待通りの結果とならない可能性がある。図9は株価情報を提供する9ヶ所のウェブサイトの提供するウェブページ1～9のそれぞれから得られた、A社の株価の現在値を例示している。このうちウェブページ4ではウェブサイトの不具合かあるいはサービス停止など何らかの理由により、文書取得部12によるウェブページの取得に失敗し情報が取得できなかったことを示している。またウェブページ2では、ウェブページそのものの取得には成功したものの、そこに含まれるA社の株価の現在値取得に失敗している。この原因としては、たとえばウェブサーバの動作不具合かウェブサイト運営者によってウェブページのデザインが変更されたことなどが考えられる。具体的には図6に示す画面表示例の、広告表示部分を削除した場合である。図6の元となるHTML文書である図5から、広告表示に相当する部分(図5(2))を削除したとすると、<TABLE>～</TABLE>で囲まれた部分は図5(3)のみとなる。すると図8に示した情報抽出方法にある「2番目の<TABLE>」は存在しないことになる。したがって、このような場合には情報要素の抽出に失敗するか、間違った情報を抽出してしまう。

【0025】またウェブページ6からは「11:00」とい

う株価を示す数値ではないデータが抽出されている。この原因としては、やはりウェブページのデザインが変更された可能性が考えられる。たとえば図5のウェブページを提供するウェブサイトが、当該ウェブページの上部に図6に示すウェブページと同様の表示を成す広告部分を追加したとする。すると図7に示す情報抽出方法における「1番目の<TABLE>」は広告表示部分を指すことになり、広告を表す部分の範囲内の文字を株価情報とみなして抽出してしまう。ウェブページ4の場合、これと同様な理由で時刻表示部分をA社の株価とみなして抽出したものと想像できる。

【0026】以上、ウェブページ2、4、6の各ケースでは株価情報が正常に抽出できなかったと判断され、最終的な出力には採用できない。したがってこれらは情報再構成部15で出力結果の対象からは外される。

【0027】次にウェブページ7からは「5,240,00」が、その他にも「¥1000」や「1000円」などの表現の異なる結果で抽出されたものがある。しかしながら後者は一般に表記され得る形の違いであり、A社の株価が1000円であることを表すウェブページが多数を占めている。よってこの図9の例からはA社の現在値の株価は1000円とするのが妥当であると判断することができる。

【0028】信頼性の高い情報要素の選定方法としては、上記のような多数決アルゴリズムが適用できる。情報源としては極力信頼のおけるウェブサイトを選定し、それぞれから目的とする情報を含むウェブページを複数取得するために位置情報を複数用意しておく。信頼性の高いウェブページばかりを取得できたとしても、既述の不具合を完全に排除することはできないが、より確からしい情報を得ることはできる。何らかの要因により正しい情報が取得できる可能性が低かったとしても、同時に複数のウェブページが同様の誤りを含むことは極めて少ないはずだからである。よって多数決の原理に従うことにより確率的に十分確からしい情報要素を推定することができる。

【0029】もっとも、信頼性の高い情報要素の方を選定できれば足りるため、多数決アルゴリズムに限定するものでもなくそれ以外の方法で行なってもかまわない。情報要素の信頼性を推定するさまざまな推論を使い分けるために、再構成手順記憶部18に推定アルゴリズムを記述するようにしてもよい。このように構成すれば、収集する情報の種類や情報源となるウェブサイトの状況に応じた適切なアルゴリズムを適用することができる。

【0030】次からはより信頼性のある情報要素の推定アルゴリズムの例について説明する。既述した多数決アルゴリズムは、信頼のおけるウェブページがある程度複数取得できなければならない。しかしながら現実には多数決アルゴリズムに必要な複数のウェブページが取得できない場合も有り得る。このような場合に有効と思われ

る推定アルゴリズムとして、a) 制約条件による方法、b) 先着の情報を優先する方法、c) 値の域値による方法などが考えられる。

【0031】a) 制約条件による方法

目的とする株価情報が2箇所のウェブページより取得できると仮定する。また、これ以外のウェブページからは取得ができないものとする。一方のウェブページではA社株価の現在値が「1,000円」、もう一方が「5,240円」という情報を提供していたとする。このような場合、どちらか一方の数値が実際に生じ得ないことが分かれば信頼性のある情報として他方の数値を採用することができる。生じ得ない情報を排斥することで、最終的な出力結果をより信頼性のあるものにすることが可能となる。

【0032】たとえば、日本の株式市場においては多くの株式に対して「値幅制限」という制限が行なわれる。図10は、この値幅制限の一例を示したものである。前日の終値が図10の左側の条件を満たす株式では、前日終値から図10の右側に示されている金額を超えて取引が行なわれないというものである。

【0033】本発明の実施例では、再構成手順記憶部18に図10の制限値幅と、これが行なわれている株式を示す情報、および同条件を逸脱する株価情報は取得した株価情報を無効とする旨の条件を記憶しておくことによって行なえる。

【0034】先の例でいえば、いまA社株式が値幅制限を受けている株式であり、A社株式の前日終値が922円であったとする。再構成手順記憶部18に記憶された上記条件を読み込んだ除法再構成部15は、この株価が値幅制限に抵触するかどうかを判断する。情報再構成部15は、抽出した情報要素がA社株式のものであることを判断し、前日終値と制限値幅表からA社株式の制限値幅が100円であることを知得する。よってA社株式の取り得る現在値は822円～1,022円の間に無ければならないことが分かる。情報再構成部15は抽出した「1,000円」および「5,240円」の値を、上記値幅に照らし合わせ「5,240円」はおおよそ有り得ない株価であると判断する。この判断を経て、情報再構成部15は「1,000円」という株価を確からしい情報要素として採用する。

【0035】上記のように目的とする情報そのものに何らかの制約があることが事前に分かっているようなものについては、再構成手順記憶部18に予め記憶しておくことによって誤った情報が混入することを避け、出力結果における情報の信頼性を高めることができる。

【0036】目的とする情報によっては制約条件を重畳することにより、さらに信頼性を高めることが可能である。たとえば株価情報の場合、1日の通算出来高は減少しない旨を制約条件として記憶しておくことにより、出来高の情報抽出における不具合を排除できる。これ以外

にも現実的には生じ得ない、あるいは極めて低い確率でしか生じ得ないと考えられる情報を検出するように条件を組み立てることも、信頼性の高い情報を構成するには有効である。先の株価情報の例でいえば、たとえば値幅制限を持たない一部の株式で、図10に示したような値幅制限の10倍を越える値幅の変動が見つかった場合、当該株価情報を採用しないというようにである。

【0037】これら情報要素の選別を行なってもなお、必要とする信頼性のある情報を取得することが困難な場合も考えられる。必要に応じて、前述した多数決アルゴリズムと併用するなど別種の判定アルゴリズムをさらに組み合わせることで効果を高めても良い。

【0038】あるいは出力結果に極めて高い信頼性を要求される場合にあっては、判定に十分な情報が収集できなかった情報に関して「取得不可」であった旨の出力をしても良い。たとえば多数決アルゴリズム併用時ににおいて取得した情報の中で全てが異なる情報を示し多数を占める値が認められない場合や、得た情報が全て制約条件範囲を逸脱するような場合である。

【0039】b) 先着の情報を優先する方法

情報提供の要請から、よりリアルタイム性を確保するために取得した情報の取得順により採用する情報要素を決定する方法である。ウェブサイトから情報を取得する場合を考えたときに、ウェブサイト毎に応答速度が異なる場合が多い。あるいは不測の事態により、特定のウェブサイトがサービスを停止している場合も有り得る。このような場合に複数の目的とする情報が収集されるのを待っている必要とするリアルタイム性を満たせないことがある。複数ある位置情報のウェブサイトから、最先に得られた目的とする情報を優先して採用する(先着優先アルゴリズム)ものである。再構成手順記憶部18が記憶する推定アルゴリズムには、最先に得られた目的とする情報を採用する旨の条件を記憶しておくことで選定がされる。

【0040】c) 値の域値による方法

取得した複数の情報のいずれかを選択する以外に、得られた複数の情報を統計的に総合して得られた結果を採用する方法も考えられる。たとえば目的とする情報が数値であった場合、複数の情報の平均値を出力とする(平均値アルゴリズム)、あるいは最大値/最小値を出力とする(最大値/最小値アルゴリズム)、などである。さまざまな統計的手法が考えられるが、情報の性質や得られる情報の活用場面を考慮して決定される。前出と同様、再構成手順記憶部18には平均/最大/最小、その他の統計的手法を用いて出力結果を構成する旨が記憶される。

【0041】今までの説明では個々に得られた複数の情報を一つの集合とみなして、その中から信頼性の高いものを推定するための方法を示してきた。次からはこれら情報を提供するウェブサイトあるいはウェブページの信

信頼性を考慮して、より信頼性の高い情報を構成する方法を説明する。

【0042】あるウェブページの取得に失敗した時、それがウェブサイトの一時的な不具合であればその復旧を待つことで、再び情報源として使用できる。しかしながらそれが一時的な要因によらない場合、たとえばウェブサービスそのものの停止やURL等の位置情報、ウェブページの構成などが予告無く変更された場合は、変更された内容によって位置情報や抽出方法の更新作業が必要となる。本来、情報源として使用することが難しいと判断される事態が判明したときは、本発明にかかる情報収集装置の管理者は当該ウェブページの取得を停止し、新たなウェブページを選定し取得先として加えるなどの管理を行なうべきである。しかしながら上記管理者が上記措置を講ずる前にも、出力として得られる情報の信頼性を維持できる仕組みが備えられていることが望ましい。

【0043】また目的とする情報の物理的な取得に失敗する以外にも、抽出した情報の一部が既述したような制約条件を満たしていないことを検出した場合には、当該位置情報から得られる目的とする情報の信頼性は全体として著しく低下したと判断すべきである。なぜなら一部に許容できない誤りがある場合、その全体の信頼性も低下していると考えられるからである。たとえばウェブサイトが提供するウェブページに広告目的で<TABLE>〜</TABLE>のような情報を付加する例を示したが、この例のように現行使用している抽出条件が適用できない危険性が極めて高い。たとえばHTMLで書かれた文書では、一部の変更の影響がHTML文書全体に及び易い。そのため抽出された情報が正しいように見えても、それは目的とする情報要素外の別の情報かも知れないからである。

【0044】このような事態を想定すると、位置情報で示されたウェブページ毎に信頼性に基づく順位を付与し、この順位に従ってそれぞれから得られる情報を差別化する必要がある。上記の順位は位置情報と共に位置情報記憶部11に記憶され、再構成手順記憶部18に記憶された手順に従って情報再構成部15によって読み出され利用される。同アルゴリズムは前出の各アルゴリズムと共に適用しても良く、そのように構成すると出力される情報の信頼性をより高めることができる。

【0045】図11は、ウェブページ1～5を提供するウェブサイトがあるときのそれぞれの抽出結果を示す図である。これは一例としての、各ウェブページの信頼性を示す順位情報に基づき情報再構成15の内部に格納されている状況を示している。上方に現れているウェブページほど信頼性が高いことを意味している。

【0046】図11ではウェブページ1および3に情報取得時に不具合が発生している。一度の情報取得失敗は、以降の情報取得によって信頼性を損なったと判断する。このため情報再構成部15は図12に示すように、各ウェブページの信頼性を示す順序を入れ替える。この

ときウェブページ3は情報抽出には成功しているものの制約条件を満たしていない情報が存在するため、最も信頼性の低い位置に入れ替えられている。先に説明したように、一部に問題を持つウェブページは全体的に問題を含んでいる可能性が高いと考えられるためである。場合によっては当該ウェブページの取得を中止するにしても良い。

【0047】ウェブページ1は取得に失敗したために信頼性の順位を下げられているが、比較的短い時間に回復が見られれば信頼性が回復され、順位が戻されるようにしても良い。このとき一向に回復が見られない（たとえば数分～数時間）ようであれば当該ウェブページのサービスの停止あるいはURLなどの位置情報が変更されたと判断し、以降の当該ウェブページの取得を中止するようにしても良い。取得を中止したウェブページは、本発明にかかる情報取得装置の管理者等が判断し必要ならば抽出条件や位置情報を更新した後に、再び信頼性を回復させる。

【0048】ウェブページ毎に信頼性を与えることができるようになると、さらに柔軟な情報取得が行なえる。たとえば各ウェブページの信頼性が図12で示される状況にあるとき、情報再構成部15は要求発行部16に指示し信頼性の高いウェブページ2、4、5のみを取得させる。このときウェブページ2、4、5から信頼性に足らない情報しか取得できなかったときに限り、ウェブページ1を取得するように構成することができる。あるいは繰り返し情報を取得するような場合には、信頼性の高いウェブページ2、4、5の取得頻度を高め、逆にウェブページ1のそれを低くするなどの差別化を図ることも可能である。

【0049】次に上述した信頼性を示す情報と、既述の多数決アルゴリズムとを併用したときの信頼性の高い情報の判定方法の例を示す。具体的には信頼性の順序が高い順にウェブページを取得し、その取得情報の中で同一のものが過半数を超えた時点で決する方法、あるいはウェブページの信頼性を数値換算し、そこから得られる情報にこの換算値を加点して行く方法とが考えられる。前者を（1）優先多数決アルゴリズム、後者を（2）議決権アルゴリズム、また両者を組み合わせたものを（3）優先議決権アルゴリズムと呼ぶ。

【0050】（1）優先多数決アルゴリズム
信頼性の判断が図12のような順序にある場合に、最上位のウェブページ（ウェブページ2）から順に取得する。情報抽出部13はそのウェブページに含まれる情報要素を抽出し、抽出した情報要素の値を蓄積する。次に2番目にあるウェブページ（ウェブページ4）から同様に情報要素を抽出し、これを蓄積する。このとき、全取得対象であるウェブページ数に占める、同じ情報要素の値が過半数を満たす数に到達するまで順次取得してゆく。例ではまず、ウェブページの2、4、5を要求し、

これらから抽出した情報要素の値が一致するならば多数決理論からその値を採用する。この時点で信頼性のある情報が確定できれば、残りのウェブページ1および3を取得する必要がない。つまりこのように判断することにより、信頼性のある情報の確定のために費やす時間を短縮できる可能性がある。仮にこの時点で過半数が得られなかった場合でも、順次つぎの信頼性を与えられているウェブページを取得し、多数決が取られた時点で終了するようにすれば良い。いずれとしても極力全てのウェブページを取得せずとも確からしい値の確定が行なえる。また信頼性のあるウェブページから取得して行くため、判定の初期から信頼性が高いと思われる情報を使用できる。このことから出力結果における信頼性を確保できる。

【0051】場合によっては多数決に必要な数のウェブページを取得できないことも考えられる。そういった場合には、適宜過半数に必要な採択数を変更しても良い。たとえばウェブページ1、4の取得に失敗したとき、ウェブページ2、4、5の3箇所から得られる情報要素のみを用いて決すれば良い。このとき過半数は2票として判定される。

【0052】取得可能なウェブページを全て取得しても、いずれの情報要素の値に対しても過半数が得られない場合も考えられる。そのような場合には、多数を得た情報要素の値とするか、取得できた内の最も信頼性の高いウェブページから得られた情報要素の値とするか、あるいはそれらの総合判断で決定すれば良い。

【0053】(2) 議決権アルゴリズム
図12に示すような信頼性順序を数値換算し、これをいわゆる議決数として多数決を取る。優先多数決アルゴリズムでは取得順に信頼性を反映させるが、本方法では投票時に重み付けを行なう点で異なる。たとえば図12の例においては信頼性順序の数値換算したものをそれぞれ「5、4、3、2、1」とすることができる。ウェブページ1〜5のそれぞれから得られる情報要素の示す値に、図12に示す順列に従ってそれぞれ「2、5、1、4、3」の議決権を割り当てる。そして情報要素の示す値の一致するものに議決権を加算し、最も高い得票数を得た情報要素の値を確からしい値として採用する。このように議決権の考え方を導入することによって、より信頼性のある値を選定することが可能となる。

【0054】(3) 優先議決権アルゴリズム
優先多数決アルゴリズムと議決権アルゴリズムを組み合わせ適用する方法である。優先多数決アルゴリズムのときと同様に、情報再構成部15は要求発行部16に指示して、信頼性の高いウェブページから順に取得していく。同時に議決権アルゴリズムのときのように、位置情報の指し示すそれぞれのウェブページに議決権を割り当てておく。そして得られた情報要素の値に一致するものに対し、当該ウェブページに割り当てられた議決権を加

算してゆく。このとき上記のように議決権を加算した後の得票数が、全得票数の過半数を獲得するに至ったかどうかを加算の都度順次判定する。特定の情報要素の値に対する得票数が過半数を越えた時点で、当該値を確からしい値として採用し、順次行なっていたウェブページの取得を終了する。

【0055】図12の例では、それぞれの信頼性順列に「5、4、3、2、1」を付与すると、全得票数は15票となる。ウェブページ2の情報要素の値が「A」であったとき、該「A」に対して5票が加算される。次のウェブページ4もまた「A」であった場合、さらに4票が加算される。このとき「A」に対する得票数は9票となり、全得票数「15票」における過半数を獲得したことになる。最短の場合、この時点で確からしい情報要素の値の判定が終了する。

【0056】上述したように、信頼性を確保したまま優先多数決アルゴリズムに比べてもより短時間で判定を終了することが可能である。

【0057】また位置情報が指し示すウェブページの信頼性を数値化し、取得の度に順列の変更が成されることにより、本発明にかかる情報収集装置の管理者にも有益である。取得の際に不具合が多く発生するウェブページは、信頼性を示す順列の下位に配置される。よって管理者は、当該ウェブページを指し示す位置情報を変更する、または情報要素の抽出方法を変更するなど、適切な措置を講ずることかできる。このことにより当該情報収集装置から得られる出力結果の信頼性を維持することが可能である。

【0058】上記のようなさまざまな方法により、より信頼性のある情報要素を選定することで、ある時点におけるより高い信頼性を持つ情報を選定し、それを再構成して出力することができる。出力の方法は、たとえば出力結果が株価情報であればグラフ化するなどして画面に表示することも可能である。あるいはファイル等に蓄積し、データとして二次利用に供することもできる。

【0059】また本発明にかかる情報収集装置では、信頼性の高い情報の選定を短時間に行なうことが可能である。これは多数決アルゴリズムのような複数の情報をつき合わせるアルゴリズムを採用すれば、統計的な計算を必要とすることもなく、あるいは一定時間経過後の前後に渡る値を採取する必要もないからである。よってリアルタイムに極めて近い情報を提供する情報収集装置を得ることも可能となる。

【0060】

【発明の効果】本発明によれば、インターネット等に一般に公開されている情報を元に、信頼性の高い情報を収集し、加えて二次利用も可能な形式の情報を得られる情報収集が可能となる。

【図面の簡単な説明】

【図1】本発明の一実施形態における情報収集装置の構

成図の一例を示す図である

【図2】本発明の一実施形態における情報収集装置の動作フローの一例を示す図である。

【図3】株価情報を提供するウェブページのHTML文書の一例を示す図である。

【図4】図3に示すHTML文書の表示イメージの一例を示す図である。

【図5】株価情報を提供するウェブページのHTML文書の一例を示す図である。

【図6】図5に示すHTML文書の表示イメージの一例を示す図である。

【図7】本発明の一実施形態における、図3に示すHTML文書から情報要素を抽出する手順の一例を示す図である。

【図8】本発明の一実施形態における、図5に示すHTML文書から情報要素を抽出する手順の一例を示す図である。

【図9】複数のウェブページから得られた情報要素の取得状況の一例を示す図である。

【図10】株式市場で採用されている、株価制限値幅表の一例を示す図である。

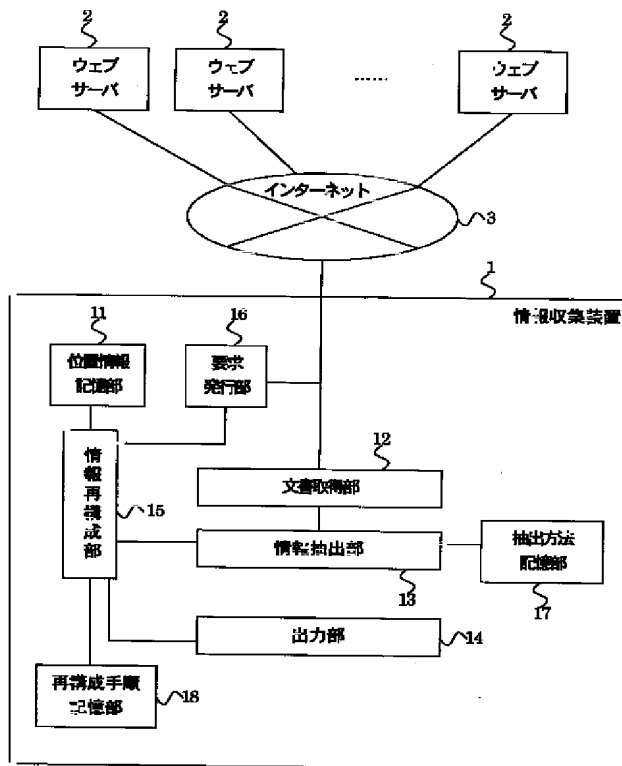
【図11】本発明の一実施形態における取得ウェブページの信頼性情報と情報抽出状況の一例を示す図である。

【図12】本発明の一実施形態における、情報抽出状況に従って並び替えた取得ウェブページの信頼性情報と情報抽出状況を示す図である。

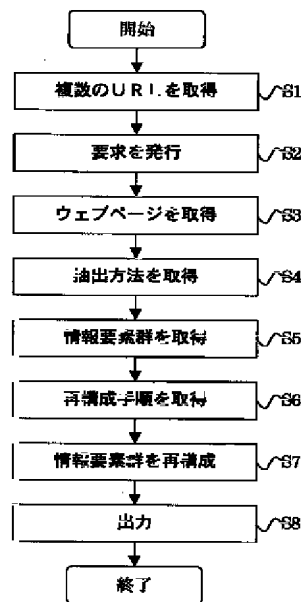
【符号の説明】

- 1 情報収集装置
- 2 ウェブサーバ
- 3 インターネット
- 11 位置情報記憶部
- 12 文書取得部
- 13 情報抽出部
- 14 出力部
- 15 情報再構成部
- 16 要求発行部
- 17 抽出方法記憶部
- 18 再構成手順記憶部

【図1】



【図2】



【図10】

前日終値	制限値幅
100 円未満	30 円
200 円未満	50 円
500 円未満	80 円
1,000 円未満	100 円
1,500 円未満	200 円
...	...

【図4】

銘柄名	始値	高値	安値	現在値	取引時刻	前日終値
A社	942	1034	930	1000	11:00	992
B社	54000	55000	49000	50000	12:35	53000
C社	49	51	49	50	10:42	49

【図3】

```

<HTML>
<HEAD><TITLE>株価情報</TITLE></HEAD>
<BODY>
<TABLE BORDER=1 CELLSPACING=0>
<TR><TH>銘柄名</TH><TH>始値</TH><TH>気値</TH><TH>安値</TH><TH>現在値</TH>
<TH>取引時刻</TH><TH>前日終値</TH></TR>
<TR><TD>A社</TD><TD>942</TD><TD>1034</TD><TD>930</TD><TD>1000</TD>
<TD>11:00</TD><TD>922</TD></TR>
<TR><TD>B社</TD><TD>54000</TD><TD>55000</TD><TD>49000</TD><TD>50000</TD>
<TD>12:35</TD><TD>56000</TD></TR>
<TR><TD>C社</TD><TD>49</TD><TD>51</TD><TD>49</TD><TD>50</TD>
<TD>10:42</TD><TD>48</TD></TR>
</TABLE>
</BODY>
</HTML>

```

【図5】

```

<HTML>
<HEAD><TITLE>現在の株価</TITLE></HEAD>
<BODY>
<TABLE>
<TR><TD><IMG SRC="images/ad1.jpg"></TD><TD><IMG SRC="images/ad2.jpg"></TD>
<TD><IMG SRC="images/ad3.jpg"></TD></TR>
</TABLE>
<TABLE BORDER=1 CELLSPACING=0>
<TR><TH>銘柄名</TH><TH>現在値</TH><TH>取引時刻</TH>
<TH>前日比</TH><TH>出来高</TH></TR>
<TR ALIGN=RIGHT><TD ALIGN=LEFT>A社
</TD><TD>1,000</TD><TD>11:00</TD><TD>78</TD><TD>5,240,000</TD></TR>
<TR ALIGN=RIGHT><TD ALIGN=LEFT>B社
</TD><TD>50,000</TD><TD>12:35</TD><TD>-6,000</TD><TD>123</TD></TR>
<TR ALIGN=RIGHT><TD ALIGN=LEFT>C社
</TD><TD>60</TD><TD>10:42</TD><TD>+1</TD><TD>8,000</TD></TR>
</TABLE>
</BODY>

```

【図6】

【図7】

広告1 広告2 広告3

銘柄名	現在値	取引時刻	前日比	出来高
A社	1,000	11:00	+78	5,240,000
B社	50,000	12:35	-6,000	123
C社	60	10:42	+1	8,000

データ開始位置	1 番目の<TABLE.>の次の</TR>の次
銘柄情報開始位置	次の<TR>の次
銘柄名	次の<TH>と</TD>の間
現在値	4 番目の<TD>と</TD>の間
取引時刻	次の<TD>と</TD>の間
前日終値	次の<TH>と</TD>の間
前日比	現在値－前日終値
データ終了位置	次の<TABLE>の位置

【図8】

【図9】

データ開始位置	2 番目の<TABLE.>の次の</TR>の次
銘柄情報開始位置	次の<TR>の次
銘柄名	次の<TD.>と</TD>の間
現在値	次の<TH>と</TD>の間のカンマ区切り数値
取引時刻	次の<TH>と</TD>の間
前日比	次の<TH>と</TD>の間のカンマ区切り符号つき数値
前日終値	現在値－前日比
データ終了位置	次の<TABLE>の位置

	A社の株価の現在値
ウェブページ1	1000
ウェブページ2	情報抽出失敗
ウェブページ3	1,000
ウェブページ4	ウェブページ取得失敗
ウェブページ5	1000 円
ウェブページ6	11:00
ウェブページ7	5,240,000
ウェブページ8	#1000
ウェブページ9	1,000 :9

【図11】

【図12】

	取得および抽出状況
ウェブページ1	ウェブページ取得失敗
ウェブページ2	正常に抽出
ウェブページ3	抽出情報の一部が制約条件を満たさない
ウェブページ4	正常に抽出
ウェブページ5	正常に抽出

	取得および抽出状況
ウェブページ2	正常に抽出
ウェブページ4	正常に抽出
ウェブページ5	正常に抽出
ウェブページ1	ウェブページ取得失敗
ウェブページ3	抽出情報の一部が制約条件を満たさない